DOI:
10.1158/2326-6066.CIR-18-0424

Document Version
Peer reviewed version

Mapping the MHC class I spliced immunopeptidome of cancer cells

Juliane Liepe[1,$], John Sidney [2], Felix K.M. Lorenz [3], Alessandro Sette [2], Michele Mishto[4,5,$]

[1] Max-Planck-Institute for Biophysical Chemistry, 37077 Göttingen, Germany

[2] Division of Vaccine Discovery, La Jolla Institute for Allergy and Immunology, La Jolla, CA 92037, United States.

[3] Max-Delbrück-Center for Molecular Medicine in the Helmholtz Association, D-13092 Berlin, Germany

[4] Centre for Inflammation Biology and Cancer Immunology (CIBCI) & Peter Gorer Department of Immunobiology, King's College London, SE1 1UL London, United Kingdom

[5] Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin, Humboldt-Universität zu Berlin, and Berlin Institute of Health, Institut für Biochemie, Germany, 10117 Berlin, Germany

[$] Correspondence: michele.mishto@kcl.ac.uk; jliepe@mpibpc.mpg.de.

Running title: MHC class I spliced immunopeptidome of cancer cells

SUMMARY (limit 38 words): Unconventional spliced peptides can be presented by cancer cells. This survey of peptide characteristics in the immunopeptidome of colon and breast carcinoma cell lines may help to predict and identify an unforeseen pool of antigenic targets for immunotherapy.

Keywords: proteasome, peptide splicing, adoptive T cell therapy targets, antigen presentation in tumors, epitopes

[$] Correspondence:
Michele Mishto,
King's College London, Faculty of Life Sciences & Medicine, School of Immunology & Microbial Sciences, Centre for Inflammation Biology and Cancer Immunology (CIBCI) & Peter Gorer Department of Immunobiology,
Great Maze Pond, 1st floor New Hunt's House, Room 1.32H, Guy's Campus, SE1 1UL London (UK).
Email: michele.mishto@kcl.ac.uk
Phone: +44(0)2078486227.

Juliane Liepe
Quantitative and Systems Biology, Max-Planck-Institute for biophysical Chemistry
Am Fassberg 11, 37077 Göttingen (D).
Email: jliepe@mpibpc.mpg.de
phone: +49(0)5512011471

The authors declare no competing financial interests.

Abstract

Anti-cancer immunotherapies demand optimal epitope targets, which could include proteasome-generated spliced peptides if tumor cells were to present them. Here, we show that spliced peptides are widely presented by MHC class I molecules of colon and breast carcinoma cell lines. The peptides derive from hot spots within antigens and enlarge the antigen coverage. Spliced peptides also represent a large number of antigens that would otherwise be neglected by patrolling T cells. These antigens tend to be long, hydrophobic, and basic. Thus, spliced peptides can be a key to identifying targets in an enlarged pool of antigens associated with cancer.

Introduction

Adoptive T cell therapy (ATT) uses $CD8^+$ T lymphocytes to selectively recognize and eliminate cancer cells. Ideal markers for cancer cell recognition are epitopes either carrying cancer-driver somatic mutations or presenting cancer-germline antigens. The number of these epitopes is, however, limited by the number of known cancer-germline genes, by the mutation frequency, and by the fact that epitopes need to have specific motifs to be presented on major histocompatibility complex class I (MHC-I) molecules and to pass all steps of the antigen presentation pathway. The identification of targetable tumor-specific epitopes is therefore one of the most challenging and yet promising quests in anti-cancer immunotherapies (1,2).

MHC-I-bound epitopes are generally produced by the proteasome, which can break proteins and release peptide fragments or re-ligate them in a process called proteasome-catalyzed peptide splicing (PCPS) (3). PCPS forms new (spliced) peptides with sequences that do not recapitulate the parental protein and is driven and regulated by factors that have been only partially determined (4-8). Despite our limited knowledge about PCPS, spliced peptides constitute a large portion of the antigenic peptide pool for human EBV-transformed cell lines and primary fibroblasts (9) and can be presented by MHC-I complexes in amounts comparable to canonical non-spliced peptides (9,10). As the non-spliced epitopes, proteasome-generated spliced epitopes can trigger *in* vivo $CD8^+$ T cell-mediated responses toward tumor-associated or pathogen-derived antigens (10-12) and can be targets for effective anti-cancer ATT (13,14).

At which magnitude could targeting spliced epitopes be an opportunity for ATTs is an open question with relevance for translational medicine. On one hand, the theoretically large variety of spliced peptide sequences suggests that recurrent driver mutations, which could not be efficiently presented on predominant MHC-I variants by non-spliced peptides because of sequence limitations, could be conversely presented by spliced peptides (15). The preliminary observation that a large portion of antigens is represented at the surface of nontumor cells only by spliced peptides (9) suggests that PCPS could permit the presentation by MHC-I molecules of so far overlooked tumor-associated antigens. On the other hand, the few examples of $CD8^+$ T cells specific for spliced epitopes described so far - derived from tumor-associated antigens (4,7,10,13,14,16) – might call into question the relevance of PCPS in generating a large number of tumor-associated epitopes. Although this question could have been already answered by studying the antigenic peptides bound to the MHC-I molecules – *i.e.* the MHC-I immunopeptidome - of cancer cell lines, technical difficulties inherent in the PCPS itself hindered that approach (15). These difficulties have been resolved through development of a strategy for the identification of spliced peptides in the MHC-I immunopeptidome by mass spectrometry (MS) (9). With that method, we could identify a portion of the immunopeptidome, which consisted of spliced peptides generated by proteasomal binding

of two peptide fragments derived from the same molecule by *cis* PCPS (**Supplementary Fig. S1**) and separated in the antigen by no more than 20 residues (9). That study, which found that an unexpectedly large frequency of spliced peptides are available for T cell recognition, demonstrated promise aid for the identification of novel targets for anti-cancer immunotherapy. Here we have developed a novel method for the identification of spliced peptides in the immunopeptidome, we have applied that approach to cancer cells and proved that PCPS enlarges the antigenic landscape in cancer cells.

Materials & Methods

Further details of the methods and the explanations of the outcomes are described in Supplemental Experimental Procedures section.

Cell lines

HCT116 and HCC1143 are cell lines derived from colon or breast carcinoma, respectively (**Supplementary Table S1**). HCC1143 have been grown in RPMI medium with 10 % FCS, 2 mM glutamin and PenStrep, 1xMEM, 1xNaPyruvat, in 5 % CO_2 atmosphere at 37° C. They have been been purchased from ATCC one years prior the use, they have been tested for mycoplasma and have not been re-authenticated.

Peptide synthesis and proteasome purification**.**

The polypeptide substrates have been synthesized using Fmoc solid phase chemistry. The sequence enumeration for the substrate polypeptides is reported in **Supplementary Table S2**. The mutated peptides (neoepitopes) identified in the MHC-I immunopeptidome of the HCT116 cell line and in the *in vitro* digestions of the synthetic substrates by purified proteasome, are reported in **Supplementary Table S3**.

20S proteasome has been purified from peripheral blood of a healthy donor, as previously described (10). Proteasome concentration has been measured by Bradford staining and verified by Coomassie staining in a SDS-PAGE gel. The purity of standardized proteasome preparations has been previously shown (17).

In vitro digestions and MS analysis**.**

Synthetic polypeptides (20 μM) have been digested by 3 μg 20S proteasome in 100 μl TEAD buffer for 20 h at 37°C as previously described (17). *In vitro* digestion samples have been measured by MS as following: 10μl digested sample has been concentrated for 5 min on a trap column (PepMap C18, 5 mm x 300 μm x 5 μm, Particel Size 100Å, Thermofisher Scientific) with 2:98 (v/v) acetonitrile/water containing 0.1% (v/v) TFA at a flow rate of 20 μl/min and then analyzed by nanoscale LC-MS/MS using an Ultimate 3000 and Q Exactive Plus mass spectrometer (ThermoFisher Scientific). The system has comprised a 75 μm i.d. × 250 mm nano LC column (Acclaim PepMap C18, 2 μm; 100 Å; ThermoFisher Scientific). The mobile phase (A) consisted of 0.1% (v/v) formic acid in water, and (B) 80:20 (v/v) acetonitrile/water containing 0.1% (v/v) formic acid. The elution has been carried out using a gradient 3-50% B in 30 min with a flow rate of 300 nl / min. Full MS spectra (*m/z* 200-2000) have been acquired on a Q Exactive at a

resolution of 70,000 (FWHM) followed by a data-dependent MS/MS of the top10 precursor ions (resolution 17,500, 4-8$^+$ charge state excluded, 1µscans). Fragment ions have been generated in a HCD cell and detected in an Orbitrap Mass Analyzer. Dynamic exclusion has been enabled with 30-s exclusion duration. The maximum ion injection time for MS scans has been set to 50 ms and for MS/MS scans to 80 ms. Background ions at *m/z* 391.2843 and 445.1200 have acted as lock mass. Peptides have been identified using the search engine Mascot version 2.6.1 (Matrix Science). The MS outcomes of the *in vitro* digestions of the synthetic substrates have been analyzed with the aim of identifying target peptides as previously described (18). Compared to the analysis method used for the analysis of the MHC-I immunopeptidomes, no restrictions for either the peptide product length or the intervening sequence length have been applied.

Extraction, processing, and analysis of proteins (> 30 kDa) from the HCC1143 cell lysate. HCC1143 cell pellet (3*10^6) has been lysed in 6 M urea/2 M thiourea in 10 mM HEPES (pH 8.0) by repeated thawing and freezing. The samples have been centrifuged at 20,000 g for 15 min at 4°C. Protein concentration in the supernatant have been quantified by BCA. Proteins larger than 30 kDa have been separated by NanoSep Centrifugal 30 kDa (Pall Life Sciences), centrifuging the sample for 15 min at 14,000 g. Then, we diluted 15 µg protein in ABC buffer (50 mM ammoniumbicarbonate, 6 mM DTT, 5% ACN). The sample was then alkylated by the addition of iodoacetamide (12 mM final concentration) and left in the dark at room temperature for 30 min. Proteins were digested by 0.3 µg Lys-C for 3 h at room temperature, further diluted in ABC buffer and digested by 0.3 µg trypsin overnight at room temperature. The sample was then purified by SepPak C18 (Waters) and eluted with a buffer 80% ACN 0.1 % TFA and concentrated by speedvac. 10µl digested sample was concentrated for 4 min on a trap column (PepMap C18, 5 mm x 300 µm x 5 µm, Particel Size 100Å, ThermoFisher Scientific) with 2:98 (v/v) acetonitrile/water containing 0.1% (v/v) TFA at a flow rate of 20 µl/min and then analyzed by nanoscale LC-MS/MS using an Ultimate 3000 and Q Exactive Plus mass spectrometer (ThermoFisher Scientific). The system comprised a 75 µm i.d. × 250 mm nano LC column (Acclaim PepMap C18, 2 µm; 100 Å; ThermoFisher Scientific). The mobile phase consisted of (A) 0.1% (v/v) formic acid in water, and (B) 80:20 (v/v) acetonitrile/water containing 0.1% (v/v) formic acid. The elution was carried out using a gradient 3-30% B in 85 min with a flow rate of 300 nl/min. Full MS spectra (*m/z* 200-2000) was acquired on a Q Exactive Orbitrap at a resolution of 70,000 (FWHM) followed by a data-dependent MS/MS of the top10 precursor ions (resolution 17,500, 4-8$^+$ charge state excluded, 1µscans). Fragment ions were generated in a HCD cell and detected in an Orbitrap Mass Analyzer. Dynamic exclusion was enabled with 20-s exclusion duration. The maximum ion injection time for MS scans was set to 50 ms and for MS/MS scans to 2000 ms. Background ions at *m/z* 391.2843 and 445.1200 acted as lock mass. Peptides were identified with the search engine Mascot version 2.6.1 (Matrix Science), applying

the same method used for the MHC-I immunopeptidome analysis (described below).

MHC-I–peptide binding affinity**.**

The binding affinity of two neoepitopes identified in the HCT116 MHC-I immunopeptidome to the four MHC-I variants (HLA-A*01:01, -A*02:01, -B*45:01, -B*18:01) of the HCT116 cell line was measured using purified MHC-I molecules, as described elsewhere (9).

Identification of spliced and non-spliced peptides

The analysis of the MS datasets, generated from the immunopeptidome of the cell lines HCC1143 and HCT116 published by Bassani-Sternberg *et al.* (19), was carried out with Mascot version 2.6.1 (Matrix Science). MS/MS scans were searched with no enzyme specificity and 6 ppm peptide precursor mass tolerance, 20 ppm MS/MS mass tolerance and HCD fragmentation.

We computed for each protein entry in the human Swissprot database all 9 to 12-mer normal and reverse *cis* spliced peptides with a maximum intervening sequence length of 25 residues (see **Supplementary Fig. S1** for the PCPS nomenclature and **Supplementary Fig. S2A,B** for the peptide identification pipeline). All spliced sequences that could be generated by simple peptide-bond hydrolysis (*i.e.* non-spliced peptides) of any human protein were removed from the database. For each resulting spliced peptide, the molecular weight (MW) was computed. Similarly, all 9-12mer non-spliced peptides and their MWs were computed. We then matched the observed precursor masses in the MS data with the MW of all theoretical spliced and non-spliced peptides that could be expected from an instrument with an accuracy of 6 ppm, and thereby reduced the overall database to a dataset-specific database. In order to generate a database that has a structure similar to that of the human proteome and to make our search strategy as similar as possible to previous studies, we transformed our spliced and non-spliced databases into the structure of the human proteome database. We concatenated the N-mer spliced peptide sequences to longer sequences, thereby generating new 'protein' entries, which have a length distribution that followed that of the human proteome. For easier annotation, we concatenated spliced and non-spliced sequences in separate 'proteins'. Based on this combined spliced and non-spliced database structure, we then computed the decoy database via randomization of the 'protein' sequences, while ensuring that none of the target N-mer spliced and non-spliced sequences is present in the decoy database. All spectra were simultaneously searched against the spliced, non-spliced and decoy database using the Mascot search engine. The search results were extracted from Mascot and filtered using an ion score cut-off, which resulted in 1% FDR. Merging short peptides into new 'proteins' results in several database entries that are artificial junctions between the short peptides; these are neither spliced peptides nor non-spliced peptides. We have considered these database entries as decoy sequences, in case they matched an MS/MS spectrum.

The protocol described so far is identical with the analysis protocol described in Liepe *et al*. (9). In order to further increase the certainty in the identification of spliced peptide sequences, we have here introduced a minimum delta score (δ) of 0.3, which describes the relative deviation of a spliced peptide ion score (s1) from a spliced or non-spliced peptide ion score (s2): $\delta = 1-s_2/s_1$, where $s_1 < s_2$. Applying this delta score avoids the annotation of MS/MS spectra with spliced peptide sequences that are not certain because very similar (spliced or non-spliced) sequences could be almost equally well matched. If $\delta < 0.3$ between two spliced peptide sequences, we did not annotated the MS/MS spectrum. If $\delta < 0.3$ between a spliced and a non-spliced peptide, we considered the non-spliced peptide as the correct assignment (given there have been no other higher scored sequences for this MS/MS spectrum). By doing so, we have made the conservative assumption that a non-spliced peptide with a score of less than 30% difference to a spliced peptide score is more likely to be the correct assignment of this MS/MS spectrum.

Since it is almost impossible to distinguish leucine (L) and isoleucine (I) from each other, we have incorporated this uncertainty in our pipeline, by checking that a spliced peptide sequence carrying either L or I could not be explained by a non-spliced peptide sequence through exchange of I and L.

We introduced a last step in the pipeline by comparing the predicted MS-HPLC retention time (or more precisely the hydrophobicity) of the peptides with the MS-HPLC retention time of the MS/MS spectrum assigned to the corresponding peptide. The hydrophobicity of peptides can be predicted depending on the MS separation system and is correlated with the measured peptide retention times. We have made use of such a predictor – *i.e.* SSRCalc, (20) - and predicted the hydrophobicity of all assigned non-spliced peptides. Assuming that the non-spliced peptide assignments are correct, we have computed the running average of the predicted hydrophobicity as a function of the measured non-spliced peptide retention time and detect the observed variance. We then applied this running average and variance to the spliced peptides. We removed all spliced peptide assignments that have predicted hydrophobicity, which shows a larger discrepancy to the running average of non-spliced peptides than the variance of the non-spliced peptides. This retention time filter is only applied to remove possibly wrong spliced peptide assignments. No non-spliced peptides were removed, regardless of their predicted hydrophobicity.

<u>Peptide sequence assignment using semi-inverted databases</u>

To estimate the possible false assignment rate of spliced and non-spliced peptide sequences in the MHC-I immunopeptidome, we generated artificial spliced peptide databases in which the sequence of either the N-terminal splice-reactant or the C-terminal splice-reactant was inverted. These semi-inverted databases contain almost the same number of sequences as the spliced human proteome database. Partially inverting sequences does not alter the molecular weight of the sequence. Therefore, the m/z-

matched semi-inverted databases contain as many entries as the m/z-matched spliced human proteome database. Since the latter two types of databases are constructed in similar ways, all database entries contain an N-mer long sequence that is identical to a peptide sequence found in the human proteome. If the identification of spliced peptides were due to an artifact appearing in the database construction and/or size, we would expect to identify similar numbers of semi-inverted peptide sequences as spliced peptides.

The semi-inverted databases were used to analyze one technical replicate (sample 20120617_EXQ0_MiBa_SA_HCT116_2_mHLA_2hr.raw) of the HCT116 immunopeptidome. The number of identified semi-inverted peptides was compared to the total number of peptides identified (sum of semi-inverted sequences and non-spliced peptides). Among the identified sequences, we checked which sequences could be spliced peptides with intervening sequences longer than 25 residues (which is the restriction we applied for the construction of the spliced human proteome database in our pipeline). In parallel, we adopted a similar strategy, in which we generated two databases where the sequences of either N-terminal or C-terminal portions of the non-spliced peptides were inverted. Since non-spliced peptides do not consist of two splice-reactants, of which we could invert one of the splice reactants, we randomly sampled an artificial peptide splicing site based on a uniform distribution along the peptide sequence. Finally, we searched the same technical replicate of the HCT116 immunopeptidome against the non-spliced proteome database together with the semi-inverted non-spliced database and considered any assignment of sequences present in the semi-inverted databases as random.

Mutations, neoepitopes and antigen presentation in the MHC-I immunopeptidome

To evaluate the prevalence of mutated antigens presented by MHC-I complexes, we calculated the number of expressed mutated and non-mutated antigens, which are represented by spliced peptides only, non-spliced peptides only, or both. The total number of expressed proteins is based on the data published by Klijin et al (21) and the mutations identified by RNAseq in independent studies (22). Further details are reported in Supplemental Experimental Procedures.

The spatial localization of the spliced and non-spliced peptides within the 3D structures of the antigens CHMP7 and RBBP7 were graphically presented using PyMol. The tertiary protein structure of CHMP7 and RBBP7 was predicted using I-Tasser (23).

Quantification

Quantification of the amount of spliced and non-spliced peptides in the immunopeptidomes by label-free MS was done as previously described (9). Briefly, we extracted the MS ion current peak area for each identified peptide (using Mascot Distiller's label-free quantification tools) and used this information to estimate the distribution of the amount of the spliced and non-spliced peptides. Potential bias of this method due to differences in the chemical features of spliced compared to non-spliced

peptides has been previously excluded (9). Further details are reported in Supplemental Experimental Procedures.

Statistical analysis.

If not described otherwise, all statistical tests have been done in R and differences in distributions have been tested using the Kolmogorov-Smirnov test. Where appropriate, p-values have been adjusted with Bonferroni correction. We computed the odds ratio (OR) of mutated antigens being represented by peptides vs. non-mutated antigens being represented by peptides by performing Fisher's exact ratio test. In this latter statistical analysis, we distinguished between all antigens represented by any non-spliced peptide and antigens represented by any spliced peptide. Correlation analysis was conducted using Pearson correlation coefficient, where the test statistic follows a t distribution.

Dataset availability

The MHC-I immunopeptidome datasets have been obtained from the PRIDE archive (identifier: PXD000394; files:
20120321_EXQ1_MiBa_SA_HCC1143_1.raw,
20120321_EXQ1_MiBa_SA_HCC1143_2.raw,
20120322_EXQ1_MiBa_SA_HCC1143_1_A.raw,
20120515_EXQ3_MiBa_SA_HCT116_mHLA-1.raw,
20120515_EXQ3_MiBa_SA_HCT116_mHLA-2.raw,
20120617_EXQ0_MiBa_SA_HCT116_1_mHLA_2hr.raw,
20120617_EXQ0_MiBa_SA_HCT116_2_mHLA_2hr.raw) or the Datadryad.org archive (doi:10.5061/dryad.r984n) and were generated by Bassani *et al.* (19), Mommen *et al.* (24). The cell source characteristics are described in **Supplementary Table S1**. The RNAseq datasets for the HCT116 and HCC1143 cell lines were obtained from Klijn *et al*. (21). The mutations' database for the HCC1143 and HCT116 refers to the Cosmic database (version 17/8/2016) (22).

All other MS files (.mgf and/or .RAW) generated for the study, the peptide spectrum matches for the immunopeptidome datasets and the mutation lists of the two cancer cell lines are available at the Mendeley database: http://dx.doi.org/10.17632/y2cvb5nvgn.1 (see **Supplementary Table S4**).

Software and computing infrastructure.

All algorithms, spliced and non-spliced peptide databases, decoy databases and data analysis and data plotting tools have been implemented in *R* on a Linux cluster system with 120 CPU-cores used for the construction of the entire human spliced peptide database (total data volume of database stored as binary RData files: 107 Gb) and for the construction of all dataset specific databases (total data volume of database stored as binary RData files per MS RAW file: 45 Gb; total data volume of database stored as FASTA file per MS RAW file: 447 Gb; Mascot compiled databases results in approximately 1.5 Tb storage space needed per RAW file analysis).

The scripts for the MHC-I spliced peptides' database generation are available at the Mendeley database: http://dx.doi.org/10.17632/y2cvb5nvgn.1.
Commercial Software: MS RAW data were converted into mascot generic file format using Mascot Distiller. Using the Mascot search engine (standard 1cpu license, uses 4 CPU-cores in parallel), the total search time per replicate of the HCC1143 and the HCT116 cell lines, respectively, was approximately 15 days (this varied depending on dataset analyzed).
The 3D representation of the 2 antigens represented by neoepitopes was carried out using the software I-Tasser (23) to predict the tertiary protein structure, and by Pymol for graphic visualization (The PyMOL Molecular Graphics System, Version 1.7.4 Schrödinger, LLC).

Results

SPI-delta *method detects MHC-I spliced immunopeptidome of cancer cell lines*

To investigate the spliced immunopeptidome of cancer cells, we developed the SPI-delta (Spliced Peptide Identifier - version delta) method (**Supplementary Fig. S2**). In particular, we allowed, in the new human spliced proteome database, spliced 9-12mer peptides generated with an intervening sequence between the splice-reactants of 25 residues or less (see **Supplementary Fig. S1** for PCPS nomenclature), allowing entries 5 residues longer than the previous version of the method (9). We introduced this modification because the study on MHC-I immunopeptidomes of nontumoral human cells did not show a prevalence of spliced peptides with short intervening sequences (9). We also introduced a minimum delta score to identify a peptide as spliced peptide. We included a final step in the pipeline to remove from the final annotation all spliced peptides that have a MS-HPLC retention time discordant to what is expected (see Materials & Methods for details and **Supplementary Fig. S2**).
We applied SPI-delta to the MHC-I immunopeptidomes of the HCT116 and HCC1143 cell lines. These cell lines were chosen because they derive from two of the most common and lethal tumors in the world (*i.e.* colon and breast cancer, respectively), and colorectal cancer can be cured by ATT by targeting neoepitopes (25).
Among the assigned sequences of the two cancer cell lines' immunopeptidomes, 1230 peptides are spliced peptides, which account for 23.6% of the variety of the immunopeptidomes (**Fig. 1A, Supplementary Table S5**). We could search the immunopeptidome samples without considering cell-specific mutations detected in these two cancer cell lines (21,22), and not allowing the identification of the most common post-translational modification (PTMs) of non-spliced peptides. In this case, the absolute number of both spliced and non-spliced peptide identifications, as well as the relative frequency of spliced peptides is increased (**Fig. 1A**). Including PTMs in the MS data analysis increases the target and therefore the decoy database, which results in more stringent cut-offs for peptide identifications to ensure 1% false discovery rate (FDR). As a

consequence, fewer peptides are identified. Ignoring PTMs can result in the assignment of MS/MS spectra as spliced peptides even though the better assignment would be a post-translationally modified non-spliced peptide. This phenomenon explains the higher frequency of spliced peptides when we do not consider PTMs.
When investigating the spliced peptide quantity, we observed that, on average, the MHC-I-restricted spliced peptides are present in smaller amounts than non-spliced peptides: spliced peptides are represented by fewer molecules than non-spliced peptides, although they represent 19.3% and 19.6% of the bulk of peptide molecules detected in the HCT116 and HCC1143 immunopeptidomes, respectively (**Fig. 1B**).
Aspects of the label-free quantification method applied to estimate the spliced and non-spliced peptide amounts are further described in Supplemental Experimental Procedures and **Supplementary Fig. S3A-G**, where further general features of the spliced peptide pool are also reported.

Validation of the cancer cell immunopeptidome assignment

One of the concerns about spliced peptide identification in the immunopeptidome is the large size of the theoretical spliced peptide database, which might result in false sequence assignments despite the strict FDR of 1% and quality control steps in the SPI-delta pipeline. To test this hypothesis, we carried out two control experiments.
In the first experiment, we generated a proteasome-independent complex peptide mixture by LysC and trypsin degradation of the HCC1143 intracellular proteome. We analyzed this dataset following the same protocol applied to the immunopeptidome, thereby using the same size of the spliced peptide database used for the immunopeptidome analysis. The sample has thousands of 9-12mer peptides and an ion charge distribution similar to the immunopeptidomes (**Fig. 1C-E**). The sample somewhat mimics the cancer cell immunopeptidome datasets. Nonetheless, only 2.4% of peptides are annotated as spliced peptides (**Fig. 1C**).
In the second experiment, as a representative example, we considered the technical replicate of the HCT116 immunopeptidome in which we identified the largest number of peptides. We analyzed it using the spliced and non-spliced human proteome databases to match the peptide precursors in the dataset (m/z matching). For those spliced peptide precursors that have been matched, we generated two databases, in which the sequence of all N-terminal or C-terminal splice-reactants have been inverted. To note, those semi-inverted databases each are approximately the size of the normal spliced peptide database generated to analyze this dataset. The HCT116 immunopeptidome dataset was then re-analyzed using the non-spliced proteome database together with either the inverted N-terminal or inverted C-terminal splice-reactant databases. Any assignment of sequences present in the semi-inverted databases is considered as randomly assigned. In parallel, we generated two databases where the sequences of all possible N-terminal or C-terminal portions (of randomly chosen length) of the non-spliced peptides have been

inverted. We then searched the same technical replicate of the HCT116 immunopeptidome against the non-spliced proteome database together with the semi-inverted non-spliced database and considered any assignment of sequences present in the semi-inverted databases as randomly assigned. These additional four searches provide us with an estimation, in the same dataset used for the MHC-I immunopeptidome identification, of the number of potentially wrongly assigned spliced peptides, depending on the database size. By searching against the semi-inverted spliced peptide databases, we assigned 4.2% C-terminally inverted spliced peptides and 3.6% N-terminally inverted spliced peptides to MS/MS spectra relative to the total number of peptides assigned (**Fig. 1F**). Some of the sequences present in the semi-inverted databases can, however, be *cis* spliced peptides with the intervening sequence longer than 25 residues. If we do not consider these latter semi-inverted peptides, 3.0% of C-terminally inverted spliced peptides and 2.5% of N-terminally inverted spliced peptides are assigned. Furthermore, we assigned 0.2% of C-terminally inverted non-spliced peptides and 1.5% of N-terminally inverted non-spliced peptides to MS/MS spectra relative to the total number of peptides assigned (**Fig. 1F**).

These results show that the number of identified spliced peptides in the cancer immunopeptidomes is higher than that of the negative controls, confirming that the spliced peptide identification in the MHC-I immunopeptidome is not an artifact due to the spliced peptide database size or structure.

Comparison of sequence motifs in MHC-I immunopeptidomes

In the MHC-I immunopeptidomes of nontumoral cells, spliced and non-spliced peptides differ in term of sequence motifs (9). This phenomenon could be due to the fact that PCPS seems to prefer sequence motifs in substrate polypeptides that are not those preferred for the normal peptide-bond hydrolysis (6). After their generation by the proteasome, peptides are subjected to sequential steps of the antigen presentation pathway that select sequence motifs (15). We would therefore expect that sequence motifs of the MHC-I-bound spliced and non-spliced peptides would cluster together because the downstream steps of the antigen presentation pathway are the same. We also expect mild differences in the sequence motifs within the clusters because of the different preferences for substrate sequence motifs of peptide hydrolysis and PCPS reactions.

Accordingly, we applied an *in silico* unsupervised approach to assign non-spliced peptides to the cancer cell lines' MHC-I variants (**Supplementary Table S1**). We identified four non-spliced peptide clusters based on their amino acid characteristics (**Fig. 2, Supplementary Fig. S4**). When we assigned each spliced peptide to the cluster possessing the most similar characteristics, we found that: (i) spliced peptides could be clustered similarly to the non-spliced peptides in both cancer cells and significantly differently to randomly assigned sequences, (ii) the resulting clusters show similar cluster

statistics (**Fig. 2A, Supplementary Fig. S4A**), and (iii) the relative distribution of spliced and non-spliced peptides in the different clusters is similar (**Fig. 2B**, **Supplementary Fig. S4B**). However, specific sequence differences emerge, despite the common overall characteristics of the grouped spliced and non-spliced peptides (see also Supplemental Experimental Procedures for additional details). In fact, the amino acid frequencies vary slightly between the same cluster of the spliced and non-spliced immunopeptidomes (**Fig. 2B**, **Supplementary Fig. S4B**), confirming our initial hypothesis. Within the same cluster, the frequency of the P1 splicing-site (see **Supplementary Fig. S1** for the nomenclature) does not always correspond to the intensity of the motif differences between spliced and non-spliced peptides. We find this unsurprising as residues around the P1 splicing-site seem to influence the PCPS efficiency (8).

PCPS enlarges the antigenic landscape of cancer cell lines

PCPS enlarges the antigenic landscape of cancer cells not only in terms of peptide variety in the immunopeptidomes but also in terms of the number of antigens presented by MHC-I-peptides. Indeed, almost 800 antigens identified in the MHC-I immunopeptidome of cancer cell lines are represented only by spliced peptides (**Fig. 3A**). These antigens could be targets for immunotherapies, if their expression is associated to the tumor or if they contain tumor-specific mutations.

The 99% and 92% of non-spliced and spliced peptides, respectively, bound to MHC-I molecules are assigned to antigens detected at transcriptional level (**Fig. 3B** and **Supplementary Table S6**; see also Supplemental Experimental Procedures for additional details). Among the 97 spliced peptides, which putatively derive from antigens not detectable in the transcriptome, 18 could derive from antigens detectable in the transcriptome if we allowed intervening sequences longer than 25 residues. Furthermore, all the remaining spliced peptides can also derive from antigens detectable in the transcriptome if we allowed PCPS between different antigens, a phenomenon called *trans* PCPS (3) that is excluded in our human spliced proteome database.

In terms of mutated antigens, 652 proteins that are identified at the RNA level in the HCT116 cell line (21) carry one or more missense mutations (22). Among them, 76% of the mutated antigens are not detected in the MHC-I immunopeptidome. The other 34%, on the contrary, is represented mainly by either spliced or non-spliced peptides, where the mutated antigens that are represented only by spliced peptides represent 5% of the mutated antigens' pool in the MHC-I immunopeptidome of the HCT116 cell line (**Fig. 3C left panel**). Both spliced and non-spliced peptides more often represent mutated antigens (131 and 51 out of 695 mutated antigens are represented by non-spliced and spliced peptides, respectively, in the merged HCC1143 and HCT116 dataset) than non-mutated antigens (2705 and 830 out of 21753 not mutated antigens are represented by non-spliced and spliced peptides, respectively, in the merged HCC1143 and HCT116

dataset; spliced peptide OR=2.0, p value=$2.34*10^{-5}$; non-spliced peptides OR=1.6, p value=$1.77*10^{-6}$). However, among those mutated antigens that are represented in the immunopeptidomes, only two mutations are actually carried by (non-spliced) peptides that have been identified in our analysis (**Fig. 3D,E** and **Supplementary Table S3**). These two neoepitopes, so named even though their recognition by T cells remains to be proved, are CHMP7[A324T]$_{316\text{-}325}$ and RBBP7[N17D]$_{12\text{-}20.}$ Both efficiently bind HLA-A*01:01 and HLA–B*18:01 molecules (**Supplementary Table S3**). Also, they can be generated by proteasome, as confirmed in *in vitro* digestions of the corresponding mutated antigenic polypeptide sequences carried out by the purified proteasome (**Supplementary Fig. S5A,B**, **Supplementary Table S2**; see also Supplemental Experimental Procedures). Transcription of both CHMP7 and RBBP7 antigens can be detected, as demonstrated in independent studies (21,22). Both antigens are represented in the MHC-I immunopeptidome by other non-spliced peptides, which, however, do not carry the mutations (**Fig. 3D,E**).

The mutation load of the HCC1143 cell line is on the contrary much smaller (21,22) and none of the mutations are carried by spliced or non-spliced peptides.

The fact that only 0.3% of the missense mutations (2 out of 695) is represented in the cancer cell lines' MHC-I immunopeptidomes confirms that tumor-specific epitopes are rare. We speculate that their identification will be facilitated by also searching for spliced peptides, even though no tumor-specific spliced peptides have been detected in these cancer cell line MHC-I immunopeptidomes (possibly due to limited sample size).

Common features of antigens in the cancer MHC-I immunopeptidomes

Regarding the antigen features in the cancer cell lines' MHC-I immunopeptidomes, the number of spliced peptides per antigen correlates with both antigen length (**Fig. 4A**) and intracellular abundance (**Fig. 4B**), as shown for non-spliced peptides (**Fig. 4A, B**), and by others (19,26,27). Furthermore, the MHC-I sampling probability (*D*), which considers the antigen length and indicates the likelihood of an antigen to be represented by a spliced peptide at the cell surface, increases for both spliced and non-spliced peptides with increasing antigen abundance (**Fig. 4C, D** and (19)). For those antigens that are represented by both spliced and non-spliced peptides, the spliced peptides' *D* correlates with the non-spliced peptides' *D* (**Fig. 4E**).

As observed also for non-spliced peptides, the likelihood of an antigen being represented in these cancer cell lines by MHC-I-spliced peptide complexes is inversely correlated with the antigen half-life. This correlation has emerged by computing the fold over representation (D/D'), where D' is the expected sampling probability, *i.e.* the average sampling probability of all antigens with the same abundance (see Supplemental Experimental Procedures for details). Indeed, the D/D' ratio inversely correlates with the antigen half-life, independently of the half-life database used in the analysis (**Fig. 4F**).

Antigen features favoring antigen coverage by spliced peptides

To identify antigenic features that result in efficient spliced peptide presentation independently of the cell types studied, we generated an extended immunopeptidome dataset by combining the datasets derived from the two cancer cell lines with that derived from the EBV-transformed lymphoblastoid cell line GR-LCL, which was generated by adopting a pre-fractioning of the peptide elution (2D strategy). The latter dataset is the most informative we have, because of the large number of identified peptides and the validation of the identifications by comparison with synthetic peptides (9). We re-analyzed the GR-LCL 2D immunopeptidome dataset by applying SPI-delta. As expected, we identified a smaller number of spliced and non-spliced peptides (**Supplementary Table S5**), thereby confirming that SPI-delta is more stringent than the previous version (9) and results in larger numbers of non-annotated MS/MS spectra.

From the extended MHC-I immunopeptidome dataset, 1096 antigens (almost 19% of all detected antigens) were identified that are represented by only spliced peptides (**Fig. 5A**). This extended immunopeptidome accounts for 11655 unique peptides, of which 9372 are non-spliced and 2283 are spliced peptides, the latter representing around 20% of the whole immunopeptidome variety. From this extended immunopeptidome dataset, we see that the presence of spliced peptides increases not only the number of antigens presented by MHC-I-peptides but also the number of peptides presented per antigen (**Fig. 5A-B**).

With this extended dataset, derived from 16 different MHC-I haplotypes (**Supplementary Table S1**), we can study the spatial distribution of spliced and non-spliced peptides by using a sliding window approach across the proteome and counting the number of observed peptides in each window (see Supplemental Experimental Procedures for details and **Supplementary Fig. S6**). We observed that spliced peptides cover a similar small fraction of the represented antigens compared to non-spliced peptides (**Fig. 5C**), although the presentation of both spliced and non-spliced peptides broadens the antigen coverage and the number of MHC-I-bound peptides per window (**Fig. 5C,D**). Furthermore, spliced peptides, like non-spliced peptides, cluster together in specific regions of the antigen, *i.e.* in "hotspots" (**Fig. 5E**). The observation that the coverage of spliced and non-spliced peptides together is smaller than the coverage of both peptide types individually (14% instead of 9.7%+7.2%=16.9%; **Fig. 5D**), indicates that they could be locally clustered together. Accordingly, the distances between spliced and non-spliced peptides are significantly smaller than the distances between randomly placed peptides (**Fig. 5E** and further details in Supplemental Experimental Procedures), hinting toward the existence of local antigenic regions prone to be represented by both spliced and non-spliced peptides.

The question, however, remains: why are some antigens represented exclusively by only spliced or only non-spliced peptides, and, what characteristics differentiate them?

Antigens represented only by non-spliced peptides, for example, are significantly shorter than those presented only by spliced peptides or by both. More hydrophobic antigens are preferentially represented by spliced peptides than non-spliced peptides. Antigens represented only by non-spliced peptides show decreasing hydrophobicity with increasing antigen length. Conversely, antigens represented only by spliced peptides have generally higher hydrophobicity than those represented only by non-spliced peptides, independently of their length, and have decreasing hydrophobicity with the increase of the length, although only until a length of approximately 1000 residues (**Fig. 6A**).
The average isoelectric point (IP) of antigens represented only by spliced peptides does not differ compared to those antigens represented only by non-spliced peptides, or represented by both types of peptides (**Fig. 6B**). However, clear differences emerge when considering the whole trimodal IP distribution - for which the average is not representative - and computing the so-called IP bias (**Fig. 6C,D**; see Supplemental Experimental Procedures for detail analysis). This latter analysis suggests that spliced peptides are generated more efficiently from basic antigens than from acidic antigens.
In summary, length, hydrophobicity and IP of an antigen are parameters that can determine whether an antigen is represented by MHC-I-spliced peptide complexes or not. The two antigens from which we have identified non-spliced peptides carrying a tumor-specific mutation, CHMP7 and RBBP7, have characteristics that favor their representation through non-spliced peptides only. They are relatively short (453 and 425 amino acids, respectively), are rather hydrophilic (hydrophobicity index of -0.48 and -0.53, respectively) and acidic (isoelectric point of 4.99 and 4.68, respectively), all characteristics disfavoring the representation by spliced peptides. Indeed, we did not detect any spliced peptide representing these two antigens.

Discussion

Despite the limited knowledge about PCPS, identification of $CD8^+$ T cells specific towards spliced epitopes and able to reduce tumor growth (13,14) hint at the value of spliced epitopes as targets for anti-cancer ATTs. This hypothesis is now supported by our demonstration here that in the breast and colon cancer cell lines, around 20% of the MHC-I immunopeptidome variety and quantity seems to be represented by spliced peptides.
This estimation depends on the technique and the statistics adopted. Indeed, the identification of spliced peptides in the immunopeptidome presents technical issues that need to be considered (15). To tackle the technical implications of the large spliced peptide database of the human proteome, we developed and applied in this study an amended identification pipeline (SPI-delta) that is stringent in term of identification confidence. For example, from the analysis of the same four MS replicates of the MHC-I immunopeptidome of the HCT116, Bassani-Sternberg and co-workers (19) identified five non-spliced neoepitopes, whereas only two of them passed our pipeline. Our results

show that the increase in stringency in our identification strategy has resulted in less sensitivity for the identification of non-spliced and spliced peptides.
Another aspect to consider is that the success in peptide identification correlates, for both spliced and non-spliced peptides, with the number of replicates analyzed. However, we found spliced peptides were less common in the cancer cell line immunopeptidomes than non-spliced peptides, as we previously observed in non-cancer immunopeptidomes (9). Therefore, the correlation is stronger for spliced peptides.
Despite our stringent pipeline, our two control experiments indicate that we still have an experimental FDR for spliced peptide identification of about 2-4%, and for non-spliced peptides of about 1%. However, none of our controls are completely free of spliced peptides. Indeed, Lys-C and trypsin can also catalyze peptide splicing (3,28). Furthermore, several among the peptides assigned as spliced peptides using the inverted splice-reactant database could be the outcome of PCPS reaction between non-contiguous peptides either with intervening sequences longer than 25 residues or derived from distinct antigens (*i.e. trans* spliced peptides). Thus, we cannot exclude the possibility of false identification for some of the peptides assigned as spliced peptides in our control experiments.
The same concept is applicable to the group of spliced peptides that, according to our mapping, are derived from antigens not detected as transcripts: several of them could be derived from antigens detected in the cancer cell transcriptome if we allowed intervening sequences longer than 25 residues and all of them could be the product of *trans* PCPS involving two antigens detected in the cancer cell transcriptome.
The exclusion of spliced peptides derived from either *trans* PCPS or *cis* PCPS with long intervening sequences, which was based on a preliminary study on one spliced epitope (29), could be misleading in the analysis of the entire MHC-I spliced immunopeptidome, as suggested by Faridi *et al*. (30). For example, we do not observe a correlation between the number of unique spliced peptides and their intervening sequence length. In the future, this restriction could be solved by adopting a de-novo sequencing strategy in the identification pipeline, for instance, as been done by Faridi *et al.* (30).
The general picture that emerges from our study points out that a large portion of the MHC-I immunopeptidome is populated by spliced peptides in cancer cell lines. Particularly relevant for anti-cancer immunotherapy could be the fact that PCPS allows representation of antigens that otherwise would be overlooked. For example, around one fourth of mutated antigens represented in the immunopeptidome of colon cancer cell line is represented only by spliced peptides, which could be relevant when searching for target neoepitopes and neoantigens suitable for ATTs. The antigens that are represented at the cell surface by spliced peptides are preferentially long, hydrophobic and basic. Why those antigens have those characteristics warrants further investigation. However, we speculate that short antigens are less likely to produce spliced peptides simply due to less combinatorial possibilities. Furthermore, spliced peptides can be produced more

easily if more hydrophobic amino acid residues are present in the antigen (8), since the C-terminal splice-reactant competes with a molecule of water for the nucleophilic attack to the acyl-enzyme intermediate (4). On the other hand, in non-spliced peptides less hydrophobic antigens are preferred, since the interaction with water molecules is needed for proteasomal peptide-bond hydrolysis. This means that the longer an antigen is, the more hydrophilic residues it needs to produce mainly non-spliced peptides. And, the longer an antigen is, the higher is the chance it would generate spliced peptides and the hydrophobicity could be progressively lower. There is, however, a hydrophobicity threshold, which is on average around 0.48 (hydrophobicity index) in our dataset, below which antigen hydrophobicity would favor the production of non-spliced peptides. Therefore, once the antigens represented only by spliced peptides reach that threshold they start to increase their average hydrophobicity.

In weighing the pros and cons of targeting spliced epitopes by anti-cancer ATTs, we shall consider that spliced peptides cluster similarly to non-spliced peptides with respect to amino acids characteristics. Some mild differences in the spliced and non-spliced sequence motifs are detectable in the cancer immunopeptidomes. On one hand this result confirms the hypothesis that spliced peptides follow the same antigen presentation pathway as non-spliced peptides and are selected by their affinity to the MHC-I cleft. On the other hand, it underlines that spliced peptides are different from non-spliced peptides. Indeed, PCPS is a very different process from peptide hydrolysis, which seems to follow different rules and to be driven by different factors (3,6,8). Only by understanding in detail those factors and dynamics, could we predict spliced peptide generation and streamline our efforts by targeting those spliced (neo)epitope candidates that most likely are efficiently produced and presented.

Acknowledgments

We thank K. Textoris-Taube and the Shared Facility Mass Spectrometry of the Charité for support in data acquisition, P. Henklein and the Peptide Synthesis Facility of the Charité for peptide synthesis. We thank D. Muharemagic for proofreading the manuscript and Prof. L. Smith for the useful discussion which helped us developing SPI-delta. The study has been in part supported by NIH to A.S. (R21AI134127), by Cancer Research UK King's Health Partners Centre at King's College London (Development Fund 2018) to MM; the experiments reported in Supplementary Fig. S5 have been performed by MM while he was appointed at Charité – Universitätsmedizin Berlin. His contract was financially supported by the Berlin Institute of Health grant awarded to P.M. Kloetzel (BIH, CRG1-TP1).

References

1. Rosenberg SA, Restifo NP. Adoptive cell transfer as personalized immunotherapy for human cancer. Science **2015**;348(6230):62-8.
2. Coulie PG, Van den Eynde BJ, van der Bruggen P, Boon T. Tumour antigens recognized by T lymphocytes: at the core of cancer immunotherapy. Nat Rev Cancer **2014**;14(2):135-46.
3. Liepe J, Ovaa H, Mishto M. Why do proteases mess up with antigen presentation by re-shuffling antigen sequences? Curr Opin Immunol **2018**;52:81-6 doi 10.1016/j.coi.2018.04.016.
4. Vigneron N, Stroobant V, Chapiro J, Ooms A, Degiovanni G, Morel S, *et al.* An antigenic peptide produced by peptide splicing in the proteasome. Science **2004**;304(5670):587-90.
5. Liepe J, Mishto M, Textoris-Taube K, Janek K, Keller C, Henklein P, *et al.* The 20S Proteasome Splicing Activity Discovered by SpliceMet. PLOS Computational Biology **2010**;6(6):e1000830.
6. Mishto M, Goede A, Taube KT, Keller C, Janek K, Henklein P, *et al.* Driving Forces of Proteasome-catalyzed Peptide Splicing in Yeast and Humans. Mol Cell Proteomics **2012**;11(10):1008-23.
7. Michaux A, Larrieu P, Stroobant V, Fonteneau JF, Jotereau F, Van den Eynde BJ, *et al.* A spliced antigenic peptide comprising a single spliced amino acid is produced in the proteasome by reverse splicing of a longer peptide fragment followed by trimming. J Immunol **2014**;192(4):1962-71.
8. Berkers CR, de Jong A, Schuurman KG, Linnemann C, Meiring HD, Janssen L, *et al.* Definition of Proteasomal Peptide Splicing Rules for High-Efficiency Spliced Peptide Presentation by MHC Class I Molecules. J Immunol **2015**;195(9):4085-95.
9. Liepe J, Marino F, Sidney J, Jeko A, Bunting DE, Sette A, *et al.* A large fraction of HLA class I ligands are proteasome-generated spliced peptides. Science **2016**;354(6310):354-8.
10. Ebstein F, Textoris-Taube K, Keller C, Golnik R, Vigneron N, Van den Eynde BJ, *et al.* Proteasomes generate spliced epitopes by two different mechanisms and as efficiently as non-spliced epitopes. Sci Rep **2016**;6:24032.
11. Platteel AC, Mishto M, Textoris-Taube K, Keller C, Liepe J, Busch DH, *et al.* CD8 T cells of Listeria monocytogenes-infected mice recognize both linear and spliced proteasome products. Eur J Immunol **2016**.
12. Platteel ACM, Liepe J, Textoris-Taube K, Keller C, Henklein P, Schalkwijk HH, *et al.* Multi-level Strategy for Identifying Proteasome-Catalyzed Spliced Epitopes Targeted by CD8+ T Cells during Bacterial Infection. Cell Rep **2017**;20(5):1242-53 doi 10.1016/j.celrep.2017.07.026.
13. Dalet A, Robbins PF, Stroobant V, Vigneron N, Li YF, El-Gamil M, *et al.* An antigenic peptide produced by reverse splicing and double asparagine deamidation. Proc Natl Acad Sci U S A **2011**;108:E323-E31.
14. Warren EH, Vigneron NJ, Gavin MA, Coulie PG, Stroobant V, Dalet A, *et al.* An antigen produced by splicing of noncontiguous peptides in the reverse order. Science **2006**;313(5792):1444-7.
15. Mishto M, Liepe J. Post-Translational Peptide Splicing and T Cell Responses. Trends Immunol **2017**;38(12):904-15 doi 10.1016/j.it.2017.07.011.
16. Hanada K, Yewdell JW, Yang JC. Immune recognition of a human renal cancer antigen through post-translational protein splicing. Nature **2004**;427(6971):252-6.
17. Mishto M, Liepe J, Textoris-Taube K, Keller C, Henklein P, Weberruss M, *et al.* Proteasome isoforms exhibit only quantitative differences in cleavage and epitope generation. Eur J Immunol **2014**;44(12):3508-21.

18. Textoris-Taube K, Keller C, Liepe J, Henklein P, Sidney J, Sette A, *et al.* The T210M Substitution in the HLA-a*02:01 gp100 Epitope Strongly Affects Overall Proteasomal Cleavage Site Usage and Antigen Processing. J Biol Chem **2015**;290(51):30417-28.
19. Bassani-Sternberg M, Pletscher-Frankild S, Jensen LJ, Mann M. Mass spectrometry of human leukocyte antigen class I peptidomes reveals strong effects of protein abundance and turnover on antigen presentation. Mol Cell Proteomics **2015**;14(3):658-73.
20. Krokhin OV, Ying S, Cortens JP, Ghosh D, Spicer V, Ens W, *et al.* Use of peptide retention time prediction for protein identification by off-line reversed-phase HPLC-MALDI MS/MS. Anal Chem **2006**;78(17):6265-9 doi 10.1021/ac060251b.
21. Klijn C, Durinck S, Stawiski EW, Haverty PM, Jiang Z, Liu H, *et al.* A comprehensive transcriptional portrait of human cancer cell lines. Nat Biotechnol **2015**;33(3):306-12 doi 10.1038/nbt.3080.
22. Forbes SA, Beare D, Bindal N, Bamford S, Ward S, Cole CG, *et al.* COSMIC: High-Resolution Cancer Genetics Using the Catalogue of Somatic Mutations in Cancer. Curr Protoc Hum Genet **2016**;91:10 1 1- 1 37 doi 10.1002/cphg.21.
23. Yang J, Yan R, Roy A, Xu D, Poisson J, Zhang Y. The I-TASSER Suite: protein structure and function prediction. Nat Methods **2015**;12(1):7-8.
24. Mommen GP, Frese CK, Meiring HD, van Gaans-van den Brink J, de Jong AP, van Els CA, *et al.* Expanding the detectable HLA peptide repertoire using electron-transfer/higher-energy collision dissociation (EThcD). Proc Natl Acad Sci U S A **2014**;111(12):4507-12.
25. Tran E, Robbins PF, Lu YC, Prickett TD, Gartner JJ, Jia L, *et al.* T-Cell Transfer Therapy Targeting Mutant KRAS in Cancer. N Engl J Med **2016**;375(23):2255-62 doi 10.1056/NEJMoa1609279.
26. Pearson H, Daouda T, Granados DP, Durette C, Bonneil E, Courcelles M, *et al.* MHC class I-associated peptides derive from selective regions of the human genome. J Clin Invest **2016**;126(12):4690-701 doi 10.1172/JCI88590.
27. Hoof I, van Baarle D, Hildebrand WH, Kesmir C. Proteome sampling by the HLA class I antigen processing pathway. PLoS Comput Biol **2012**;8(5):e1002517 doi 10.1371/journal.pcbi.1002517.
28. Berkers CR, de Jong A, Ovaa H, Rodenko B. Transpeptidation and reverse proteolysis and their consequences for immunity. Int J Biochem Cell Biol **2009**;41(1):66-71.
29. Dalet A, Vigneron N, Stroobant V, Hanada K, Van den Eynde BJ. Splicing of distant Peptide fragments occurs in the proteasome by transpeptidation and produces the spliced antigenic peptide derived from fibroblast growth factor-5. J Immunol **2010**;184(6):3016-24.
30. Faridi P, Li C, Ramarathinam SH, Vivian JP, Illing PT, Mifsud NA, *et al.* A subset of HLA-I peptides are not genomically templated: Evidence for cis- and trans-spliced peptide ligands. Sci Immunol **2018**;3(28) doi 10.1126/sciimmunol.aar3947.
31. Boisvert FM, Ahmad Y, Gierlinski M, Charriere F, Lamont D, Scott M, *et al.* A quantitative spatial proteomics analysis of proteome turnover in human cells. Mol Cell Proteomics **2012**;11(3):M111 011429 doi 10.1074/mcp.M111.011429.
32. McShane E, Sin C, Zauber H, Wells JN, Donnelly N, Wang X, *et al.* Kinetic Analysis of Protein Stability Reveals Age-Dependent Degradation. Cell **2016**;167(3):803-15 e21 doi 10.1016/j.cell.2016.09.015.

Figure legends

Figure 1. Size of the MHC-I spliced and non-spliced immunopeptidomes of colon and breast carcinoma cell lines and related controls. (**A**) Number of spliced and non-spliced peptides identified in the MHC-I immunopeptidomes of HCT116 and HCC1143 cell lines, as well as their combined dataset. In the upper panel, we report the absolute and relative frequency of spliced and non-spliced peptides considering the mutations detected in the two cell lines and several PTMs. The latters were allowed only for the non-spliced peptide database. However, we report here the absolute and relative frequency of non-spliced peptides not post-translationally modified to be comparable to spliced peptides (also not carrying PTM). In the lower panel, we report the absolute and relative frequency of spliced and non-spliced peptides not considering the mutations detected in the two cell lines and PTMs. (**B**) Distribution of the MS ion peak area of spliced and non-spliced peptides in the HCT116 and HCC1143 immunopeptidomes measured by label-free quantification. The MS ion peak area distribution of the non-spliced peptides is significantly larger than the distribution of spliced peptides in both immunopeptidomes (Kolmogorov-Smirnov test; HCC1143 p-value: 0.00156; HCT116 p-value: 0.03). The total abundance of spliced peptides calculated from the integral of the MS ion peak areas of spliced peptides relative to the integral of the peak area of all peptides is reported. The number of identified peptides and the MS ion peak area correlates significantly with the number of biological replicates in which they are identified (see **Supplementary Fig. S3B-E**). (**C**) Frequency of spliced and non-spliced peptides detected in the LysC-trypsin digestion of the HCC1143 intracellular proteasome-unprocessed proteome. Proteins larger than 30kDa have been separated from the cell lysate of the HCC1143 cell line - to eliminate protein fragments already produced by proteasome – and digested by LysC and trypsin. The resulting sample has been analyzed by MS using the same data analysis strategy used for the MHC-I immunopeptidomes. Shown are the number and frequencies of spliced and non-spliced peptides assigned in the LysC/trypsin-processed proteome dataset (n=1). (**D**) Comparison of peptide length distribution identified in the HCC1143 immunopeptidome (left panel) and the LysC/trypsin-processed intracellular proteome (right panel). (**E**) Comparison of the detected precursor charge distribution of the HCC1143 immunopeptidome and the LysC/trypsin-processed intracellular proteome. (**F**) Frequency of detected semi-inverted spliced peptides (green/ yellow) compared to frequency of non-spliced and spliced target peptides identified in one technical replicate of the HCT116 immunopeptidome. Indicated percentages are relative to the total number of peptides assigned in each experiment. The frequency of identification of semi-inverted spliced peptides is an estimation of the frequency of wrongly annotated spliced and non-spliced peptide sequences. Yellow indicated fractions are semi-inverted peptide sequences, which could also be explained

as target *cis* spliced peptides with intervening sequence length longer than 25 residues (see Materials & Methods).

Figure 2. Sequence motifs of the MHC-I spliced and non-spliced immunopeptidome of the colon carcinoma cell line. (**A**) Distribution of the distances within the cluster of non-spliced peptides (orange line), between spliced and non-spliced peptides (blue line) and between non-spliced peptides and control random peptides (grey line) in the 4 clusters of the spliced and non-spliced peptides identified in the MHC-I immunopeptidome of the HCT116 cell line. The lower panel shows the Kolmogorov-Smirnov distance between the distributions of non-spliced and spliced peptides and control peptides, respectively, which are significantly different (Kolmogorov-Smirnov test; p-value = 0.03). (**B**) Comparison of the amino acid frequencies for each position of the non-spliced and spliced 9mer peptides of the HCT116 MHC-I immunopeptidome, after clustering according to their amino acid features. For each of the 4 clusters, amino acid frequencies are shown in the left panels. The size of the amino acid letters corresponds to their occurrence within the cluster. The number of peptides belonging to each cluster and their relative frequency are also reported. On the right panels, the motifs' difference between the amino acid frequencies of the non-spliced and spliced 9mer peptides is reported as Jensen-Shannon (JS) divergence. The inlets on the top of the right panels show the frequency of PCPS (as P1 position) for each residue. The HLA-I alleles corresponding to each cluster are reported, and they have been identified by similarities with known HLA-I-specific peptide sequence motifs.

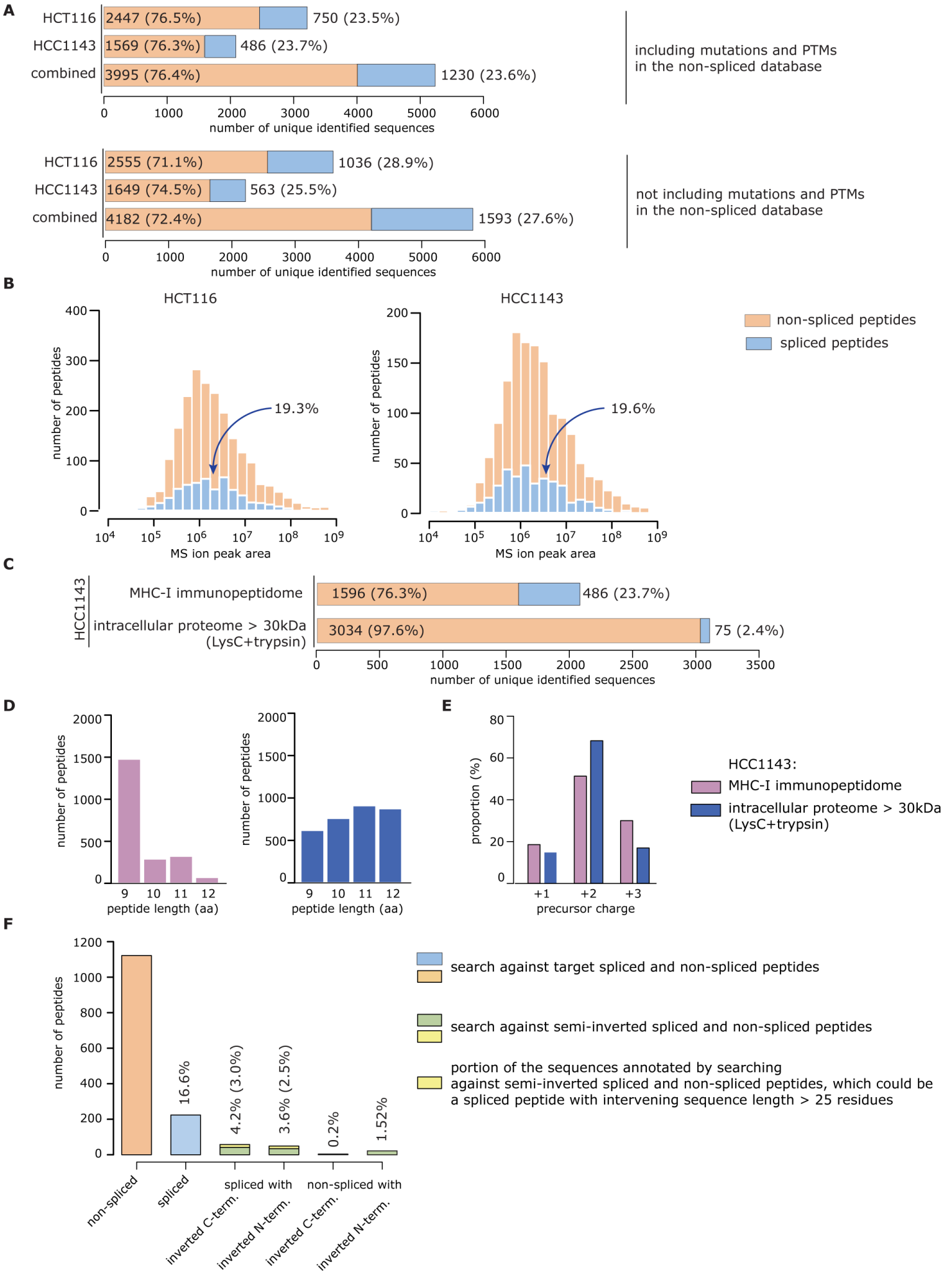
Figure 3. PCPS enlarges the antigenic landscape of the two cancer cell lines. Data refer to the peptides identified in the MHC-I immunopeptidomes of the HCT116 and HCC1143 cell lines, which is reported as a combined analysis of the two datasets in (**A,B**). (**A**) Number of antigens represented by only spliced peptides, by only non-spliced peptides, or by both in the immunopeptidome. (**B**) Number and frequency of sequences that are present in the immunopeptidomes and that derive from antigens also detected in the cancer cell line transcriptomes (21), considering separately antigens represented by spliced or non-spliced peptides. (**C**) Prevalence of HCT116 and HCC1143 mutated proteins represented (or not represented) on MHC-I complexes by either spliced peptides, non-spliced peptides, or both. The frequencies refer to either proteins that were detected at the RNA level and have missense mutations (left panels) or proteins that were detected at RNA level regardless their mutational load (right panels). (**D,E**) 3D structures of the antigens CHMP7 (**D**), and RBBP7 (**E**), from which 5 non-spliced peptides were identified in the HCT116 immunopeptidome, including the two non-spliced neoepitopes CHMP7[A324T]$_{316\text{-}325}$ and RBBP7[N17D]$_{12\text{-}20}$. The structures were predicted using i-Tasser server. The non-spliced peptides (including the non-spliced neoepitopes) are labeled in orange.

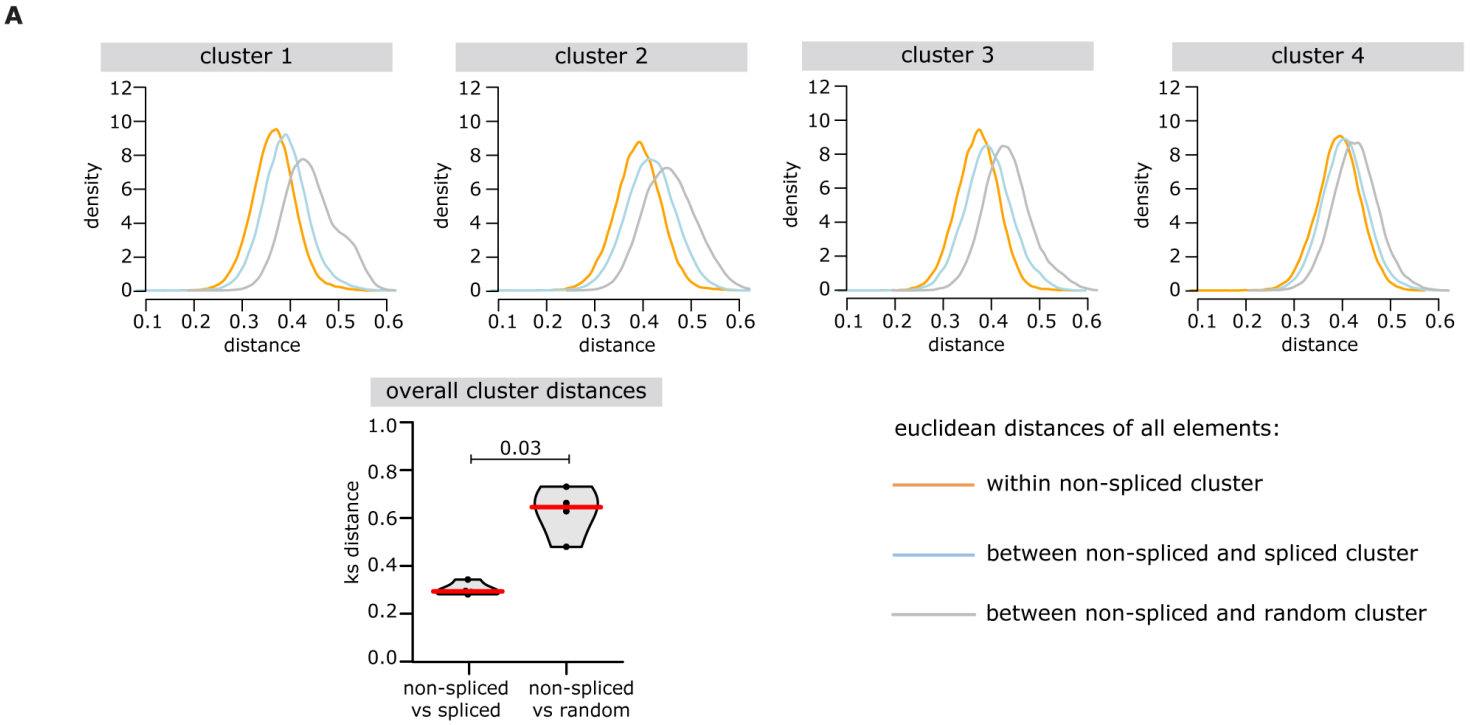
Figure 4. Relationship between antigen length, abundance and half-life detected in the cancer cell lines and their probability of representation as MHC-I-spliced peptides. Data refer to the antigenic peptides identified in the MHC-I immunopeptidomes of the HCT116 and HCC1143 (here reported as a combined analysis of the two datasets). (**A**) Correlation between the number of spliced or non-spliced peptides detected in the immunopeptidomes and the antigen length (non-spliced peptides: C=0.15, p-value< 10^{-16}; spliced peptides: C=0.15, p-value< 10^{-16}). (**B**) Correlation between the number of spliced (light blue dots) or non-spliced (orange dots) peptides per antigen and the antigen abundance as measured by Bassani-Sternberg *et al*. (19) in the cell lysates. Dark blue lines and red lines indicate a running average of the peptide numbers over the antigen abundance. Both the number of spliced and non-spliced peptides is correlated with antigen abundance (non-spliced: C = 0.7, p value < 10^{-16}; spliced peptides: C = 0.4, p = < 10^{-16}). In (**A,B**) the Y-axis is in log scale. (**C**) Relationship between the number of theoretically possible 9mer spliced or non-spliced peptides, respectively, and the antigen length. The number of non-spliced peptides has been computed as N = antigen length – 8 (red line). The number of unique spliced peptides in the human proteome spliced peptide database was counted and could be approximated by linear regression (dark blue line) with n = 399.12*antigen length – 7948.4 (p-value < 10^{-16} for both estimated parameters using linear regression in R). The dashed green line indicates the ratio of the number of spliced peptides over the number of non-spliced peptides, which asymptotically reaches 398 for antigens longer than 500 amino acids. (**D**) Correlation between the MHC-I peptide sampling probability (*D*) and the antigen intensity measured in the intracellular proteome (19). (**E**) Correlation between the spliced peptide sampling density and the non-spliced peptide sampling density (C = 0.7, p value < 10^{-16}). (**F**) Relationship between the fold over representation (D/D') and the antigen half-life, using half-lives based on Boisvert *et al*. (31), or McShane *et al*. (32). Only antigens identified in the intracellular proteome of the HCT116 and HCC1143 cell lines by Bassani-Sternberg *et al.* (19) have been included here.

Figure 5. Spliced peptides broaden the antigens' coverage of tumor and nontumor cell lines and locally cluster with non-spliced peptides in antigen hot spots. The values refer to the extended human MHC-I self-immunopeptidome, which includes the immunopeptidomes of HCT116 and HCC1143 cancer cell lines, and GR-LCL. (**A**) Number of antigens represented by only spliced peptides, only non-spliced peptides, or both. Among all identified antigens, 1096 antigens are represented by 1197 unique spliced peptides, 3850 antigens are represented by 6987 non-spliced peptides and 910 antigens are represented by both spliced (n=1095) and non-spliced (n=2481) peptides. (**B**) Frequency of non-spliced peptides, spliced peptides, or any peptide per antigen. (**C**) Coverage of the antigen sequences by either non-spliced peptides, spliced peptides, or

both considering a 25 or 50 residue window, respectively. (**D**) Distribution of the number of antigenic peptides (non-spliced, spliced or both) per window (using a 50-residue window). Percentage represents antigen coverage. (**E**) Measured distance between either non-spliced peptides, spliced peptides, or between spliced and non-spliced peptides. The red lines represent the respective random distributions, for which no local clustering can be observed and which significantly differs from the distribution of the distance of the peptides identified in the MHC-I immunopeptidomes (Mann-Whitney test p-values are shown).

Figure 6. Antigens represented by either spliced or non-spliced peptides have different characteristics. The values refer to the extended human MHC-I self-immunopeptidome, which includes the immunopeptidomes of HCT116 and HCC1143 cancer cell lines, and GR-LCL. (**A**) Correlation between the average hydrophobicity and length for the antigens represented by non-spliced peptides (tan line), spliced peptides (blue line), or both (grey line), and antigens not represented in the extended MHC-I self-immunopeptidome (black line). Running averages are shown. (**B**) Correlation between the average IP and length for antigens represented by non-spliced peptides (tan line), spliced peptides (blue line), or both (grey line, and antigens not represented in the extended MHC-I self-immunopeptidome (black line). Running averages are shown. (**C**) Distributions of computed IPs for all antigens represented by either non-spliced peptides, spliced peptides, or both. All three groups show trimodal distributions (grey histograms), which could be approximated as a Gaussian mixture model (black line) consisting of three Gaussian distributions with differing mean and standard deviations (red, yellow and green lines). The two Gaussian distributions with average IP < 7 (red and yellow lines) include the acidic set of antigens, whereas the Gaussian distribution with average IP > 7 (green lines) include the set of basic antigens. (**D**) IP-bias (isoelectric point bias) for antigens represented by either non-spliced peptides, spliced peptides, or both. The IP-bias is the proportion of antigens that fall into the basic set compared to the acidic set (color scheme corresponds to **C**).


A
HCT116
2447 (76.5%)
750 (23.5%)
HCC1143
1569 (76.3%)
486 (23.7%)
combined
3995 (76.4%)
1230 (23.6%)
0 1000 2000 3000 4000 5000 6000
number of unique identified sequences
including mutations and PTMs in the non-spliced database
HCT116
2555 (71.1%)
1036 (28.9%)
HCC1143
1649 (74.5%)
563 (25.5%)
combined
4182 (72.4%)
1593 (27.6%)
0 1000 2000 3000 4000 5000 6000
number of unique identified sequences
not including mutations and PTMs in the non-spliced database
B
HCT116
HCC1143
number of peptides
19.3%
19.6%
MS ion peak area
non-spliced peptides
spliced peptides
C
HCC1143
MHC-I immunopeptidome
1596 (76.3%)
486 (23.7%)
intracellular proteome > 30kDa (LysC+trypsin)
3034 (97.6%)
75 (2.4%)
0 500 1000 1500 2000 2500 3000 3500
number of unique identified sequences
D
number of peptides
peptide length (aa)
E
proportion (%)
precursor charge
HCC1143:
MHC-I immunopeptidome
intracellular proteome > 30kDa (LysC+trypsin)
F
number of peptides
16.6%
4.2% (3.0%)
3.6% (2.5%)
0.2%
1.52%
non-spliced
spliced
spliced with inverted C-term.
spliced with inverted N-term.
non-spliced with inverted C-term.
non-spliced with inverted N-term.
search against target spliced and non-spliced peptides
search against semi-inverted spliced and non-spliced peptides
portion of the sequences annotated by searching against semi-inverted spliced and non-spliced peptides, which could be a spliced peptide with intervening sequence length > 25 residues


A
cluster 1
cluster 2
cluster 3
cluster 4
density
distance
overall cluster distances
ks distance
0.03
non-spliced vs spliced
non-spliced vs random
euclidean distances of all elements:
within non-spliced cluster
between non-spliced and spliced cluster
between non-spliced and random cluster

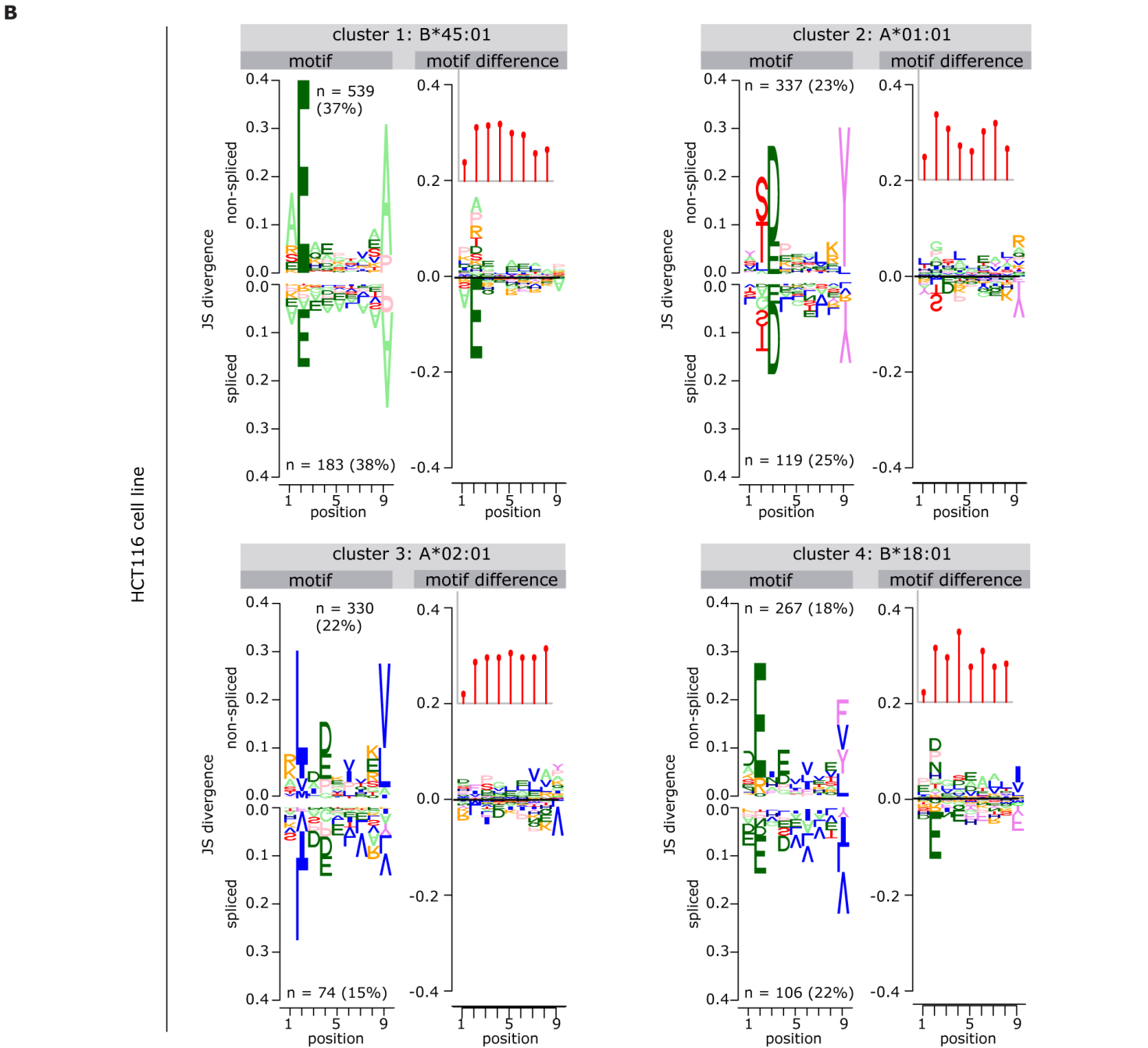

B
HCT116 cell line
cluster 1: B*45:01
cluster 2: A*01:01
cluster 3: A*02:01
cluster 4: B*18:01
motif
motif difference
n = 539 (37%)
n = 183 (38%)
n = 337 (23%)
n = 119 (25%)
n = 330 (22%)
n = 74 (15%)
n = 267 (18%)
n = 106 (22%)
non-spliced
spliced
JS divergence
position

Figure 3

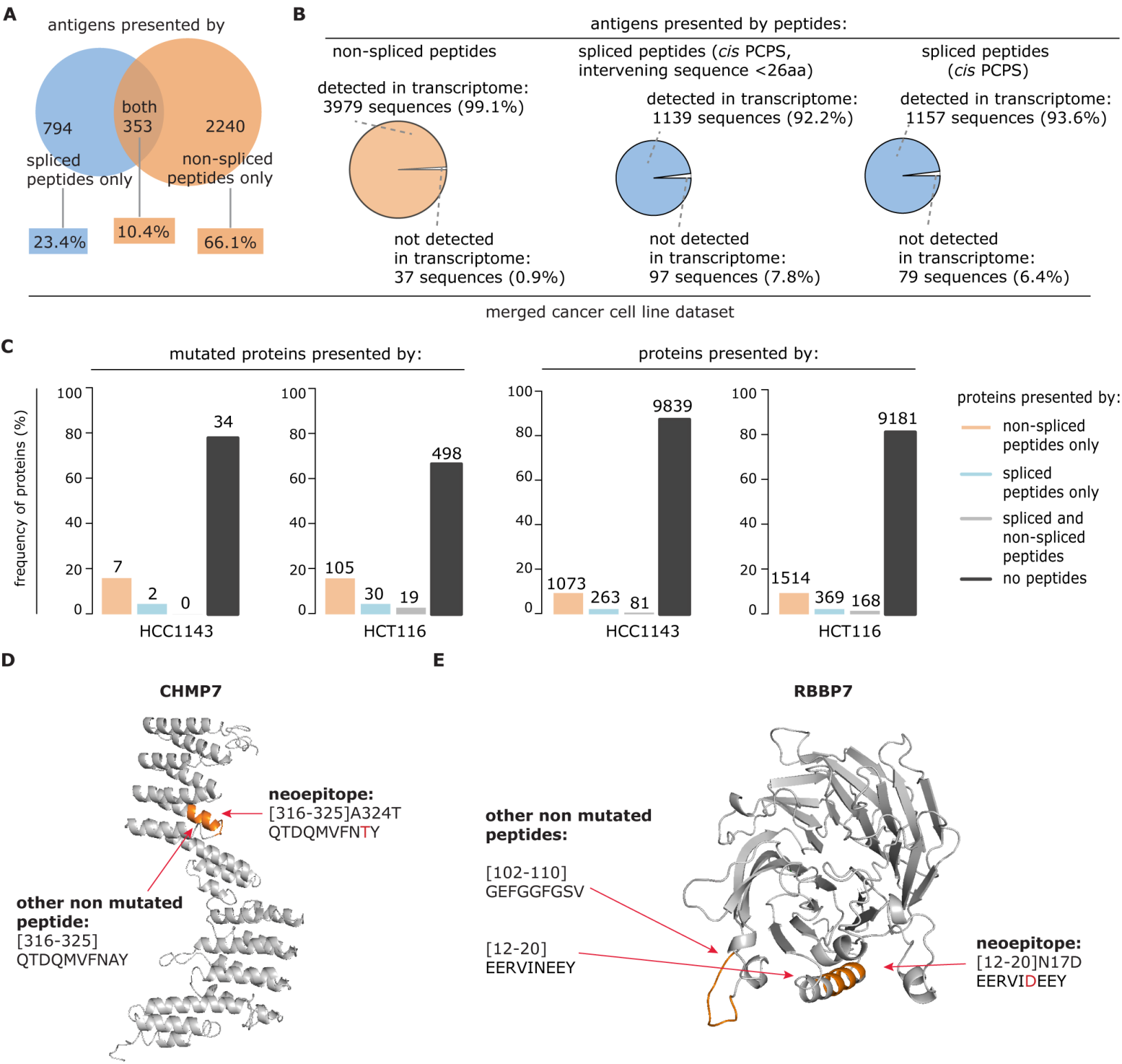

A
antigens presented by
794
both
353
2240
spliced
peptides only
non-spliced
peptides only
23.4%
10.4%
66.1%
B
antigens presented by peptides:
non-spliced peptides
detected in transcriptome:
3979 sequences (99.1%)
not detected
in transcriptome:
37 sequences (0.9%)
spliced peptides (cis PCPS,
intervening sequence <26aa)
detected in transcriptome:
1139 sequences (92.2%)
not detected
in transcriptome:
97 sequences (7.8%)
spliced peptides
(cis PCPS)
detected in transcriptome:
1157 sequences (93.6%)
not detected
in transcriptome:
79 sequences (6.4%)
merged cancer cell line dataset
C
mutated proteins presented by:
proteins presented by:
frequency of proteins (%)
HCC1143
HCT116
proteins presented by:
non-spliced peptides only
spliced peptides only
spliced and non-spliced peptides
no peptides
D
CHMP7
neoepitope:
[316-325]A324T
QTDQMVFNTY
other non mutated peptide:
[316-325]
QTDQMVFNAY
E
RBBP7
other non mutated peptides:
[102-110]
GEFGGFGSV
[12-20]
EERVINEEY
neoepitope:
[12-20]N17D
EERVIDEEY

Figure 4

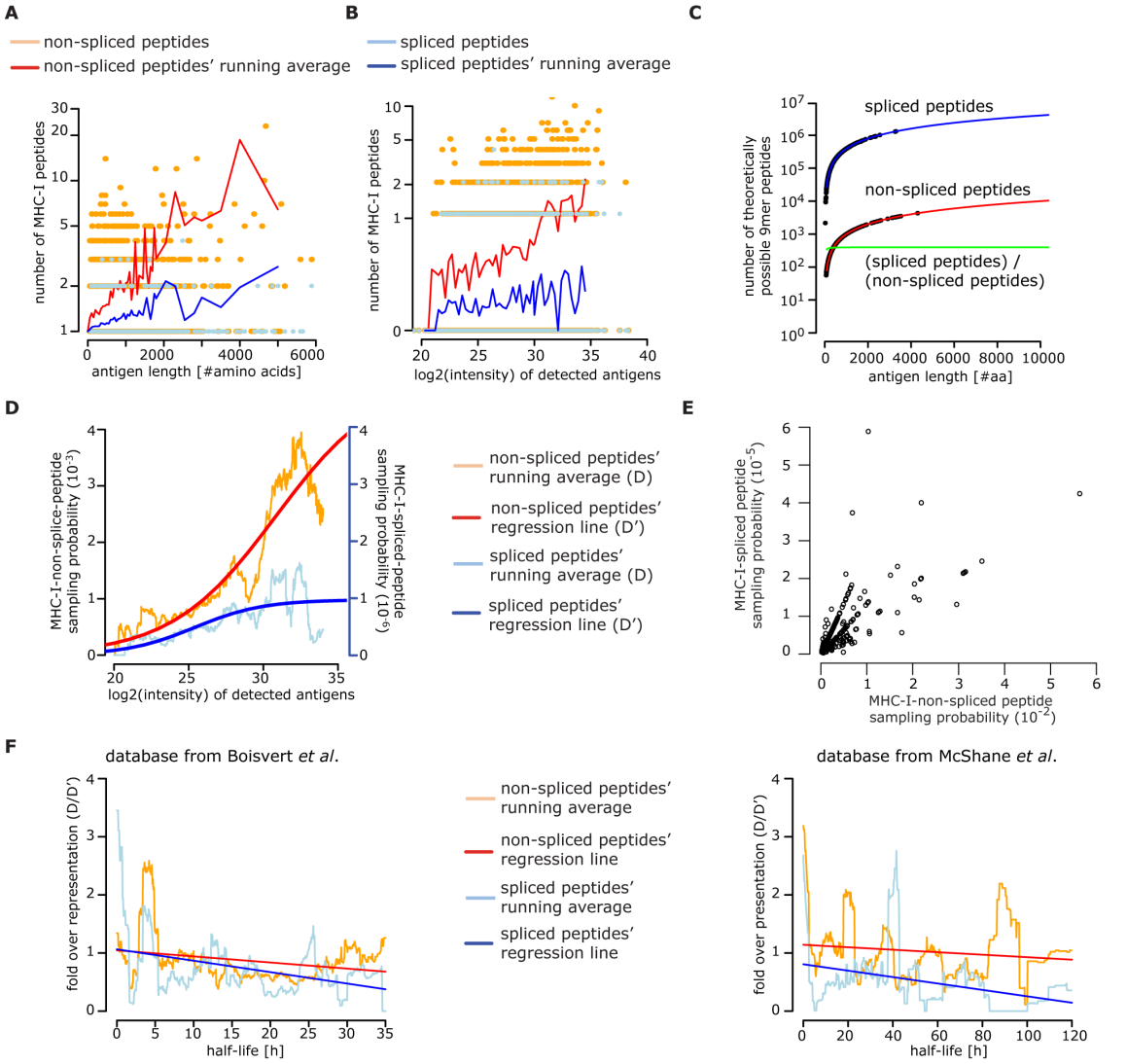

A
non-spliced peptides
non-spliced peptides' running average
number of MHC-I peptides
antigen length [#amino acids]
B
spliced peptides
spliced peptides' running average
number of MHC-I peptides
log2(intensity) of detected antigens
C
spliced peptides
non-spliced peptides
(spliced peptides) / (non-spliced peptides)
number of theoretically possible 9mer peptides
antigen length [#aa]
D
MHC-I-non-splice-peptide sampling probability (10^-3)
MHC-I-spliced-peptide sampling probability (10^-5)
log2(intensity) of detected antigens
non-spliced peptides' running average (D)
non-spliced peptides' regression line (D')
spliced peptides' running average (D)
spliced peptides' regression line (D')
E
MHC-I-spliced peptide sampling probability (10^-5)
MHC-I-non-spliced peptide sampling probability (10^-2)
F
database from Boisvert et al.
fold over representation (D/D')
half-life [h]
non-spliced peptides' running average
non-spliced peptides' regression line
spliced peptides' running average
spliced peptides' regression line
database from McShane et al.
fold over presentation (D/D')
half-life [h]

Figure 5

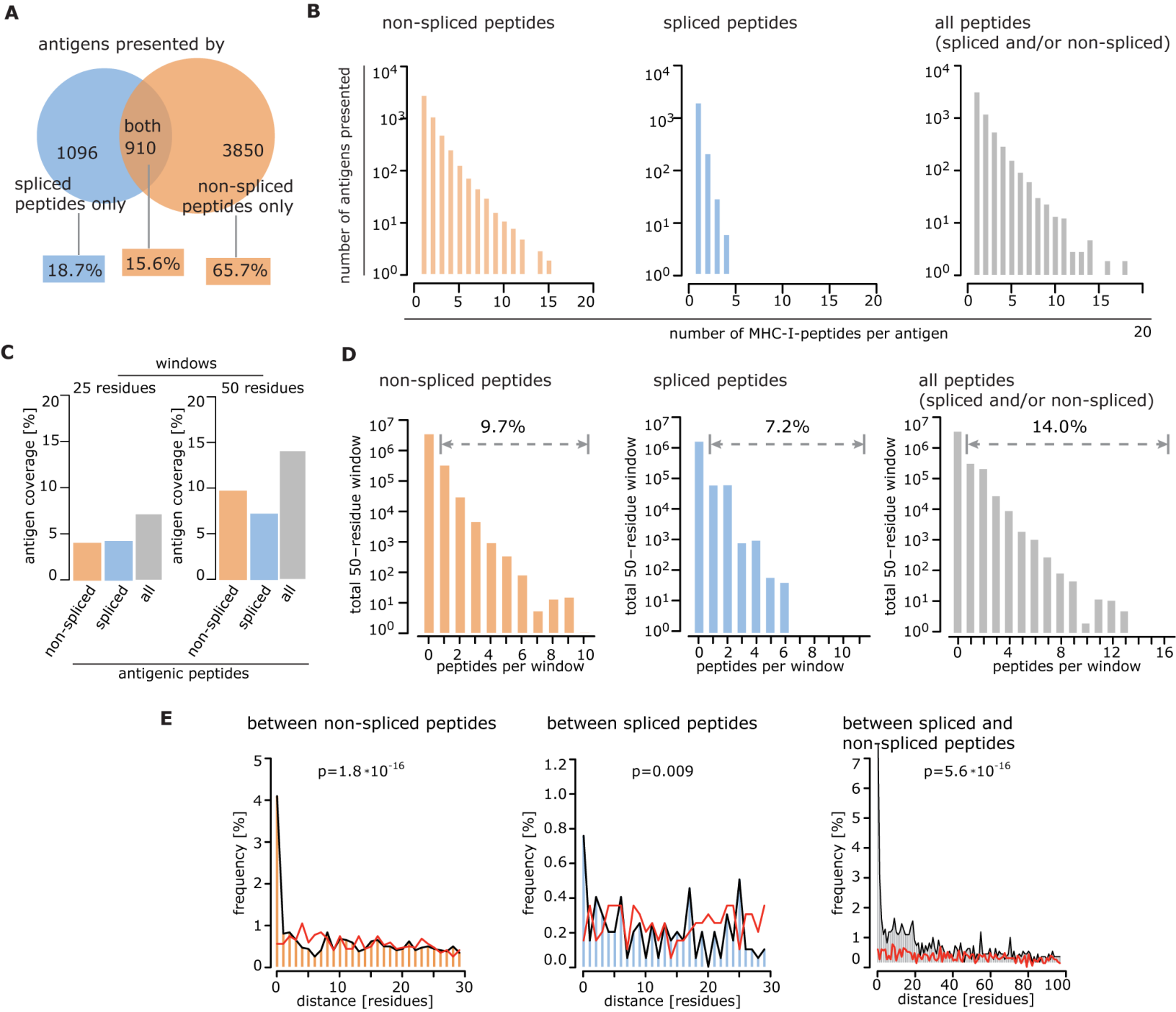
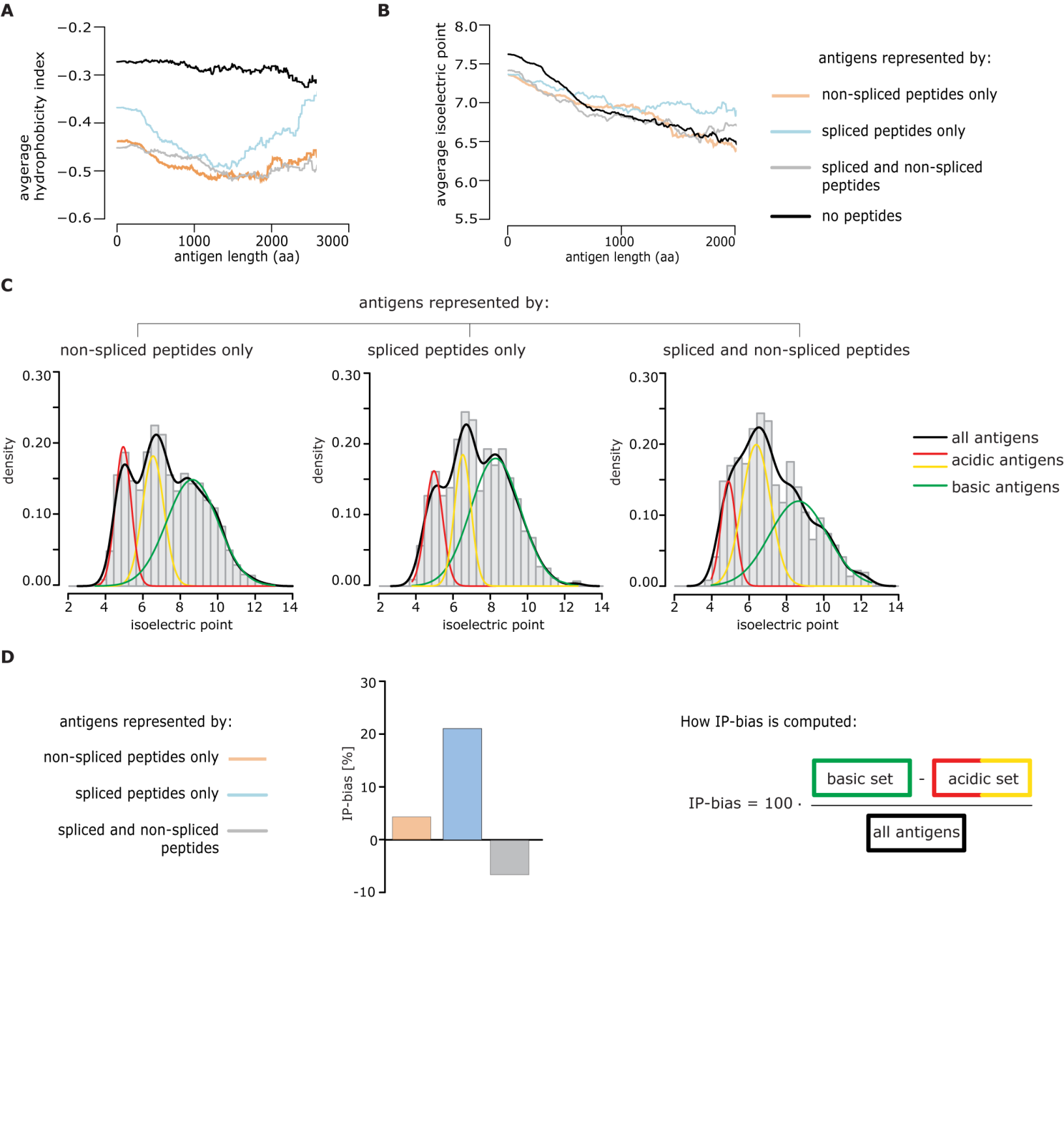

A
antigens presented by
1096
both
910
3850
spliced
peptides only
non-spliced
peptides only
18.7%
15.6%
65.7%
B
non-spliced peptides
spliced peptides
all peptides
(spliced and/or non-spliced)
number of antigens presented
number of MHC-I-peptides per antigen
C
windows
25 residues
50 residues
antigen coverage [%]
non-spliced
spliced
all
antigenic peptides
D
non-spliced peptides
spliced peptides
all peptides
(spliced and/or non-spliced)
9.7%
7.2%
14.0%
total 50−residue window
peptides per window
E
between non-spliced peptides
between spliced peptides
between spliced and
non-spliced peptides
p=1.8*10-16
p=0.009
p=5.6*10-16
frequency [%]
distance [residues]

Figure 6


A
avgerage hydrophobicity index
−0.2
−0.3
−0.4
−0.5
−0.6
0
1000
2000
3000
antigen length (aa)
B
avgerage isoelectric point
8.0
7.5
7.0
6.5
6.0
5.5
0
1000
2000
antigen length (aa)
antigens represented by:
non-spliced peptides only
spliced peptides only
spliced and non-spliced peptides
no peptides
C
antigens represented by:
non-spliced peptides only
spliced peptides only
spliced and non-spliced peptides
density
0.30
0.20
0.10
0.00
2
4
6
8
10
12
14
isoelectric point
all antigens
acidic antigens
basic antigens
D
antigens represented by:
non-spliced peptides only
spliced peptides only
spliced and non-spliced peptides
IP-bias [%]
30
20
10
0
-10
How IP-bias is computed:
IP-bias = 100 ·
basic set
-
acidic set
all antigens